

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problems Mailbox.**

2/9/1

DIALOG(R) File 347:JAPIO
(c) 2000 JPO & JAPIO. All rts. reserv.

04629821 **Image available**
WHOLE SENTENCE DATA BASE RETRIEVING METHOD

PUB. NO.: 06-301721 JP 6301721 A]
PUBLISHED: October 28, 1994 (19941028)
INVENTOR(s): OTA HIROSHI
NOMURA YASUTSUGU
KUREBAYASHI TETSUYA
YAMADA KIYOSHI
NEGISHI MASAMITSU
OYAMA KEIZO
MIYAZAWA AKIRA
SHIBANO KOJI
APPLICANT(s): HITACHI LTD [000510].(A Japanese Company or Corporation), JP
(Japan)
APPL. NO.: 05-091080 [JP 9391080]
FILED: April 19, 1993 (19930419)
INTL CLASS: [5] G06F-015/40
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)
JOURNAL: Section: , Section No. FFFFFFF, Vol. 94, No. 10, Pg. FFFFFF,
FF, FFFF (FFFFFF)

ABSTRACT

PURPOSE: To provide the whole sentence data base retrieving system by which the designation of a retrieval request sentence containing the logical structure information of a document and the display of a result of retrieval can be executed, in the processing for retrieving the whole sentence of the document having a logical structure.

CONSTITUTION: An original document described by an inputted SGML is decomposed, based on its structure information, a data base of structure information 7, tag information 8 and text information 9 is generated by a data base generating part 3 of an SGML server, the generated data base is stored in a data base control system, a retrieval request 16 inputted by the SGML is converted to an SQL being data base retrieval language and a retrieval processing is executed by a relational data base control system, and a result of retrieval 15 is organized to an SGML 17, and displayed on a work station.

?t 3/9/1

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平6-301721

(43)公開日 平成6年(1994)10月28日

(51)Int.Cl.⁵
G 0 6 F 15/40

識別記号 庁内整理番号
5 0 0 D 9194-5L

F I

技術表示箇所

審査請求 未請求 請求項の数1 O L (全10頁)

(21)出願番号 特願平5-91080

(22)出願日 平成5年(1993)4月19日

(71)出願人 000005108

株式会社日立製作所
東京都千代田区神田駿河台四丁目6番地

(72)発明者 太田 浩史

東京都江東区新砂一丁目6番27号 株式会
社日立製作所公共情報事業部内

(72)発明者 野村 泰嗣

東京都江東区新砂一丁目6番27号 株式会
社日立製作所公共情報事業部内

(72)発明者 紅林 徹也

東京都江東区新砂一丁目6番27号 株式会
社日立製作所公共情報事業部内

(74)代理人 弁理士 小川 勝男

最終頁に続く

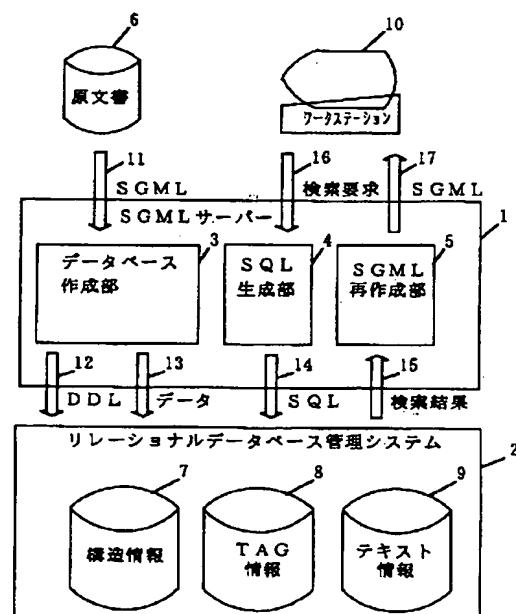
(54)【発明の名称】 全文データベース検索方法

(57)【要約】

【目的】論理的構造をもった文書の全文を検索する処理において、文書の論理的構造情報を含めた検索要求文の指定及び検索結果の表示を可能とする全文データベース検索方式を提供することにある。

【構成】入力されるSGMLで記述される原文書をその構造情報をもとに分解し、SGMLサーバのデータベース作成部3で構造情報7、タグ情報8、テキスト情報9のデータベースを作成し、データベース管理システムに作成するデータベースを格納し、SGMLで入力される検索要求16をデータベース検索言語であるSQL14に変換してリレーショナルデータベース管理システムで検索処理を行い、検索結果15をSGML17に編成し、ワークステーションに表示する。

図1



【特許請求の範囲】

【請求項1】論理的構造をもった文書の全文を検索する方法において、前記文書の構成要素単位で文書を分解して、文書の論理的構造を示す構造情報と、前記構成要素を識別するための構成要素識別子と、前記構成要素識別子に続くテキスト情報を含むデータベースを作成するデータベース作成ステップと、入力された前記構成要素識別子を含む検索要求文をデータベース検索用の言語記述に変換する検索要求文生成ステップと、前記検索要求文生成ステップにより生成される検索要求文に基づいて前記データベースを検索するステップと、前記検索ステップにより得られた検索結果を前記構成要素識別子とこれに続くテキスト情報の形式に変換する検索結果編成ステップと、前記検索結果編成ステップにより編成された検索結果を表示する表示ステップとを設けることを特徴とする全文データベース検索方法。

【発明の詳細な説明】**【0001】**

【産業上の利用分野】本発明は、予め人手によるキーワードの作成を必要とせずに文章に含まれる全ての単語を検索対象とすることが可能である全文データベース検索方法、特に論理的構造をもった文書データをデータベースとして格納し検索するデータベース検索方法に関する。

【0002】

【従来の技術】従来の文書検索方式では、キーワード作成に必要とする工数およびキーワードの付け方が悪かったことによるヒット漏れの問題を解決する方法として、文献全文を検索対象とする全文データベース検索方式が用いられている。しかし、この全文データベース検索方式では全文書が検索範囲となり、無駄なヒットが生じる。この問題を解決する手段として、文書の論理構造を指定して検索要求を行ない、検索効率を上げる方法が用いられている。このような検索を行う検索方式として、文書を何らかの方法で分解し蓄積して、検索処理を行うものがある。このような検索方式の公知例として、例えば、特開平01-189721号公報、特開平01-189722号公報では、文書を予め設定された複数の軸により整理して文書を格納し、検索時にはそれらの軸により検索を指定する方式がある。しかし、この手法では文書構造を指定した検索処理は行えない。また、特開平03-62169号公報では、文書を画像、図形、テキストの情報に分離し各メディアごとにデータの構造を自動的に抽出し分解し、マルチメディアファイルとして格納し検索を行う。

【0003】

【発明が解決しようとする課題】上記従来技術では、文書の分解を画像、図形、テキストなどのメディアごとに、行い、マルチメディア情報としてデータベースに格納するために、画像、図形に対する検索要求を実行する

ことが可能である反面、専用のマルチメディアデータベース管理システムを開発する必要がある。

【0004】本発明の目的は、論理的構造をもった文書をその構成要素単位に分解し、汎用のリレーションナルデータベースに格納することにより、データベース利用者が認識している文書構造と検索システムが認識しているデータ構造を一致させ、その文書構造情報を含む検索要求をリレーションナルデータベース検索用の記述言語で記載することにより、文書の構造情報を含む検索要求文の指定による検索結果の表示を可能とする全文データベース検索方法を提供することにある。

【0005】

【課題を解決するための手段】本発明は、文書の構成要素単位で文書を分解して、文書の論理的構造を示す構造情報と、この構成要素を識別するための構成要素識別子例えばタグと、この構成要素識別子に続くテキスト情報を含むデータベースを作成するデータベース作成ステップと、入力された構成要素識別子を含む検索要求文をデータベース検索用の言語記述に変換する検索要求文生成ステップと、検索要求文生成ステップにより生成される検索要求文に基づいてデータベースを検索するステップと、得られた検索結果を構成要素識別子とこれに続くテキスト情報の形式に変換する検索結果編成ステップと、この検索結果を表示する表示ステップとを設ける。

【0006】

【作用】論理的構造をもった文章の全文データベース検索処理において、データベース作成ステップでは入力される原文書を分割し、文書の構造情報と、その構成要素を識別するための構成要素識別子と、構成要素識別子に続くテキストとを含むデータベースを作成し、検索要求文生成ステップでは検索処理の際に指定される検索要求文を文書の論理的構造とデータベース構造の対応関係をもとにデータベース検索用の言語記述に変換し、検索結果編成ステップでは検索処理の結果得られる文書データをそのID、順序番号をもとにソートして結合することにより文書の構造情報を含む言語記述に変換し、表示ステップでは表示装置に検索結果を表示することにより、文書の構造情報を含む検索要求文の指定と、検索結果の表示を可能とする。

【0007】

【実施例】以下、本発明の一実施例について図面を用いて説明する。

【0008】図1は本発明の一実施例のシステム構成図である。本実施例では、原文書はSGMLで記載され、データベース管理システムとのインターフェースにはSQLを用いるものとする。SGML(Standard Generalized Markup Language)とは、ISO(国際標準機構)で定められた規格で、文書を構造化することを可能とする汎用組版指示言語である。SQL(Structured Query

Language) はリレーションナルデータベースの標準処理言語である。SGMLを用いた文書の構造の表現にはタグが用いられる。タグには開始タグと終了タグがあり、文書構造情報の構成要素を開始タグと終了タグで囲むことにより、文書中の文章の区切りと、その文章が文書構造上どの構成要素に属するのかを明確にする。ここで開始タグは要素名称を記号「<」と「>」で閉じたものであり、終了タグは要素名称を「</」と「>」で閉じたものである。タグに続く構成要素の内容がテキストである。検索対象である構造情報を含む文書は、タグとテキストからなる。

【0009】図1に基づいて本発明を実現するシステム構成を説明する。本発明は、SGMLサーバ1および汎用のリレーションナルデータベース管理システム2により構成され、さらにSGMLサーバ1はデータベース作成部3、SQL生成部4、SGML再作成部5により構成される。データベース作成部3は文書構造情報を内部に持つ言語の1つであるSGML11で記載された原文書を入力し、入力されたSGMLに基づいてリレーションナルデータベース管理システム2にデータベースのテーブル定義を行うための言語であるDDL(Data Definition Language)12を発行することにより、SGMLによる構造情報7、タグ情報8、テキスト情報9のテーブルを作成し、SGMLを用いた原文書データのデータ13を格納する。SQL生成部4ではワークステーション10より文書構造を指定可能な検索言語DQL(Document Query Language)で作成される検索要求16を入力し、SQL14を用いた検索要求文を作成し、リレーションナルデータベース管理システム2に対して検索要求を発行する。SGML再作成部5では、リレーションナルデータベース管理システム2でSQL生成部4により生成され、SQL14を用いた検索要求文により検索された検索結果15について構造情報7を用いてタグを含むSGML17を編成し、ワークステーション10に表示する。

【0010】図2(a)は、図1のSGML11で表現される文書の論理構造のモデルであり、一般誌101が表紙102より段落109までの8つの構成要素により構成されている。図2(b)は、SGMLで表現される文書の構成要素の性質を示す図の一例である。

【0011】図3は論理構造を持つSGMLで表現される文書の一例であり、タグ201より208を用いてテキスト211より214の論理構造が表現されている例である。

【0012】図4から図6は、図2、図3のSGMLで表現される文書をリレーションナルデータベースに格納する場合の各テーブルデータフォーマットの一例である。図4の構造情報テーブルのデータフォーマットは図2(a)、(b)のモデルをテーブル化したもので、上下の構造を各要素のタグ302にタグID301を付加す

るとともにその上位タグID303と関係付け、同一階層にある構成要素間の定義される順序は順序番号304として、その定義の順序に従って表し、タグが出現するか否かは連結タイプ305として表わす。

【0013】図5のタグ情報テーブルのデータフォーマットは図4のタグの階層毎に作成するものであり、図3のSGMLのタグごとにIDを付加し、その上位レベルで出現したタグのタグIDとともに格納される。同一文書で同一種のタグに対して複数回出現する可能性のあるタグに対しては著者タグテーブル504の著者NO.423のようにNO.を付加する。

【0014】図6のテキスト情報テーブルのデータフォーマットはSGMLで記載された原文書から分解されたテキストをその要素別にそのタグID及び文書中での一連の通し番号であるテキストNO.を付加して格納するものである。

【0015】図4から図6のIDの付け方は一例であり、キーとしてユニークなものであれば良い。

【0016】図7は本発明によるデータベース検索処理の流れで、図8は本発明によるデータベース検索処理の実行例である。ワークステーション10からDQLに従って入力された検索要求は、SQL生成部4で処理される。例えば、「章に「印刷」を含む一般誌の表紙」という検索要求が、DQL701としてワークステーション10より入力される。DQL701のFROM句712より図2の文書構造情報を呼出し(601)、DQL701のWHERE句713の「章」は構成要素として章106、表題107、節108、段落109を持ち、テキストを持つ要素は表題107、節108、段落109であることが分かる。テキストを持つ図6の表題テキスト513、節テキスト514、段落テキスト515のそれぞれのテーブルに対してWHERE句713の「LIKE'%印刷%'」を展開し(602)、表題ID417、節ID418、段落ID419を求め、さらにそれぞれのIDが図5の表題タグ507、節タグ508、段落タグ509の其れ其れのテーブルで一致する上位構造の要素である章IDを求めるというようにWHERE句713で指定した構成要素の章IDが得られるまで図5のテーブルに検索を行い、さらにそれら全てを組み合わせるSQLを用いた検索要求文を作成する。図2の文書構造情報から、SELECT句711とWHERE句713で指定される「表紙」と「章」の要素間の階層の上下関係がないと判定し(603)、図2の文書構造情報から表紙102と章106の最も下位レベルで共通の上位要素として存在する一般誌101が最終的に求める要素IDと決定され(605)、図5の章タグ506のテーブルで章IDが一致する本文ID415を得るというよう、一般誌IDが得られるまで図5のテーブルに検索を行うように繰返し展開し(607)、最終結果を一時テーブルに格納する。指定された検索要求をSQLで

記載された検索要求文に変換し、SQLで記載された検索要求文702を完成させ検索を実行する。SELECT句711で指定されている要素「表紙」は、図2の文章構造情報から表紙102-題名103、表紙102-著者104という2通りの経路でテキストを持つ構造にたどりつくことが分かり、処理607で求められた一時テーブルの一般誌ID101の下位に存在する全要素のIDとテキストの順序番号と内容を得られるようなSQL731とSQL732をそれぞれの経路で作成して実行し、検索結果741と検索結果742を得る(608)。検索結果741と検索結果742の表紙IDが一致していることと順序番号から、同一の表紙に属し、題名「よくわかるDQL」、著者「日立太郎」の順で並んでいることを解析する(609)。処理610で処理609で解析されたテキストにタグを付加し、SGMLに従った検索結果文705を編成する(610)。

【0017】

【発明の効果】本発明によれば以下のような効果が得られる。

【0018】論理的構造をもった文書を検索する処理において、文書をその構成要素単位に分解し、データベースに格納し、それを検索し、検索結果を構成要素識別子とテキスト情報の形式に変換し表示することにより、特定の言語を用いる論理的構造情報を含む検索要求文の指定と、構造情報を含む検索結果の表示が可能となる。

【図面の簡単な説明】

【図1】本発明の一実施例のシステム構成を示す図である。

【図2】論理的文書構造のモデルを示す図及び論理的構造情報である文書の種類の一例を示す図である。

【図3】SGML文書の一例を示す図である。

【図4】リレーションナルデータベース管理システムに格納する構造情報テーブルの一例である。

【図5】リレーションナルデータベース管理システムに格納する構造情報記号テーブルの一例を示す図である。

【図6】リレーションナルデータベース管理システムに格納するテキスト情報テーブルの一例を示す図である。

【図7】本発明における検索処理の流れの一例を示す図である。

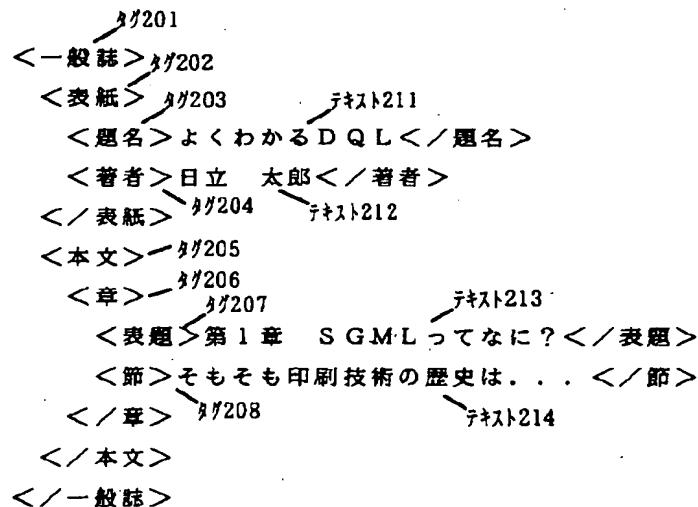
【図8】本発明における検索処理実行の一例を示す図である。

【符号の説明】

- 1 SGMLサーバ
- 2 リレーションナルデータベース管理システム
- 3 データベース作成部
- 4 検索要求文生成部
- 5 検索結果編成部
- 6 原文書
- 7 構造情報のデータベース
- 8 タグ情報のデータベース
- 9 テキスト情報のデータベース
- 10 検索用ワークステーション

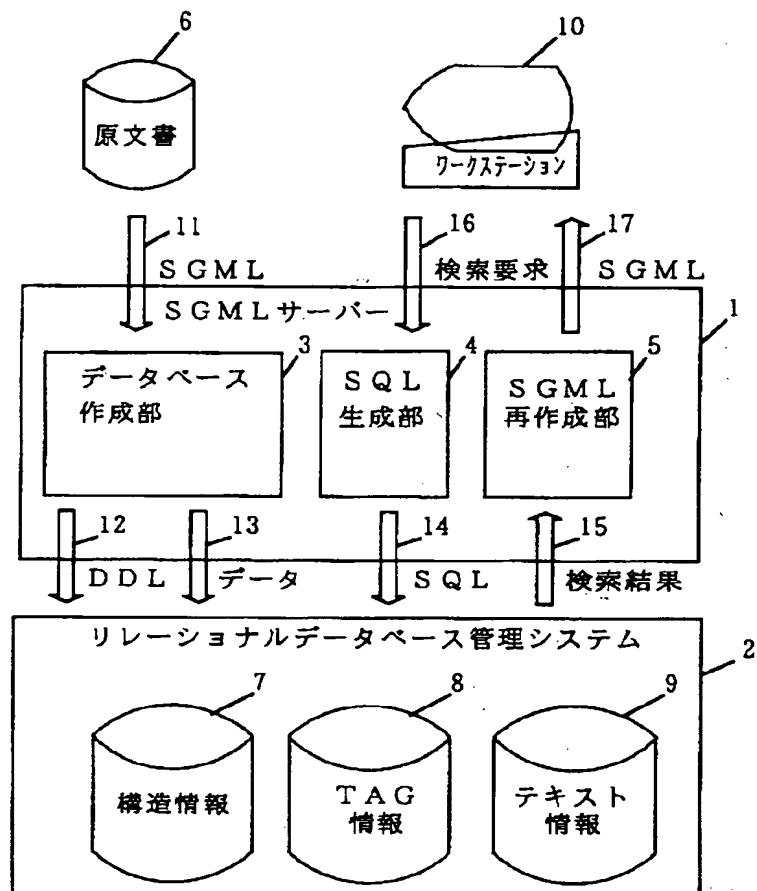
【図3】

図3



【図1】

図1



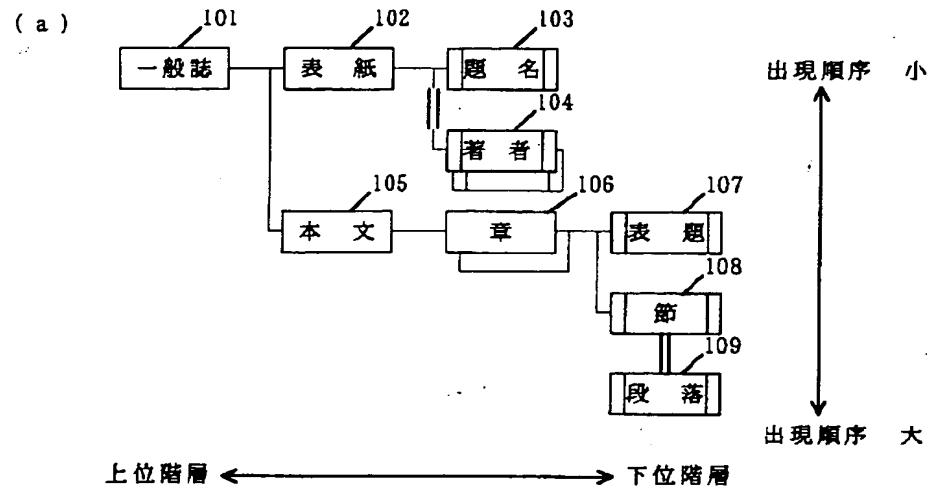
【図4】

図4

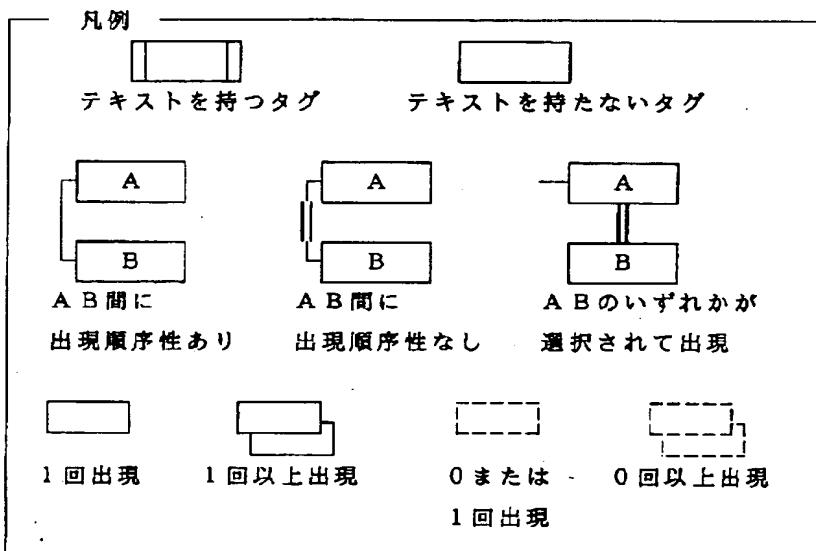
タグ I D	タグ名	上位タグ I D	順序番号	連結タイプ	305
0 0 1	一般 誌	101	1	順序あり	
0 0 2	表紙	102 0 0 1	1	順序あり	
0 0 3	題名	103 0 0 2	1	順序なし	
0 0 4	著者	104 0 0 2	1	順序なし	
0 0 5	本文	105 0 0 1	2	順序あり	
0 0 6	章	106 0 0 5	1	順序あり	
0 0 7	表題	107 0 0 6	1	順序あり	
0 0 8	節	108 0 0 6	2	選択出現	
0 0 9	段落	109 0 0 6	2	選択出現	

【図2】

図2



(b)



【図5】

図5

表紙タグ 502 411 一般誌 I D 表紙 I D 412 001001 002001	題名タグ 503 412 413 表紙 I D 題名 I D 002001 003001
著者タグ 504 412 413 423 表紙 I D 著者 I D 著者 N O 002001 004001 1	本文タグ 505 411 415 一般誌 I D 本文 I D 001001 005001
章タグ 506 415 416 426 本文 I D 章 I D 章 N O 005001 006001 1	表題タグ 507 416 417 章 I D 表題 I D 006001 007001
節タグ 508 416 418 章 I D 節 I D 006001 008001	段落タグ 509 416 419 章 I D 段落 I D

【図6】

図6

題名テキスト 511

題名 I D 003001	テキスト N O 1	テキスト よくわかる D Q L	413	431	432	211
------------------	---------------	---------------------	-----	-----	-----	-----

著者テキスト 512

著者 I D 004001	テキスト N O 2	テキスト 日立 太郎	414	431	432	212
------------------	---------------	---------------	-----	-----	-----	-----

表題テキスト 513

表題 I D 007001	テキスト N O 3	テキスト 第1章 S G M Lってなに？	417	431	432	213
------------------	---------------	--------------------------	-----	-----	-----	-----

節テキスト 514

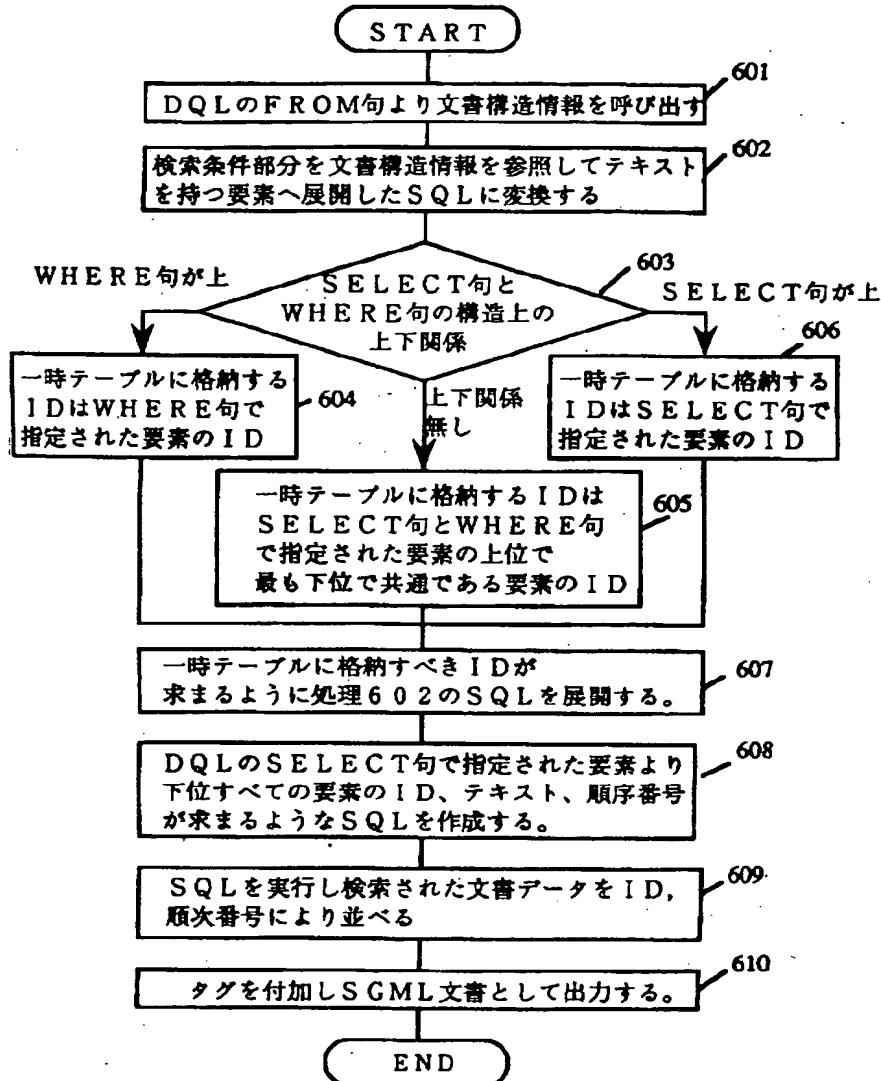
節 I D 008001	テキスト N O 4	テキスト そもそも印刷技術の歴史は...	418	431	432	214
-----------------	---------------	-------------------------	-----	-----	-----	-----

段落テキスト 515

段落 I D	テキスト N O.	テキスト	419	431	432

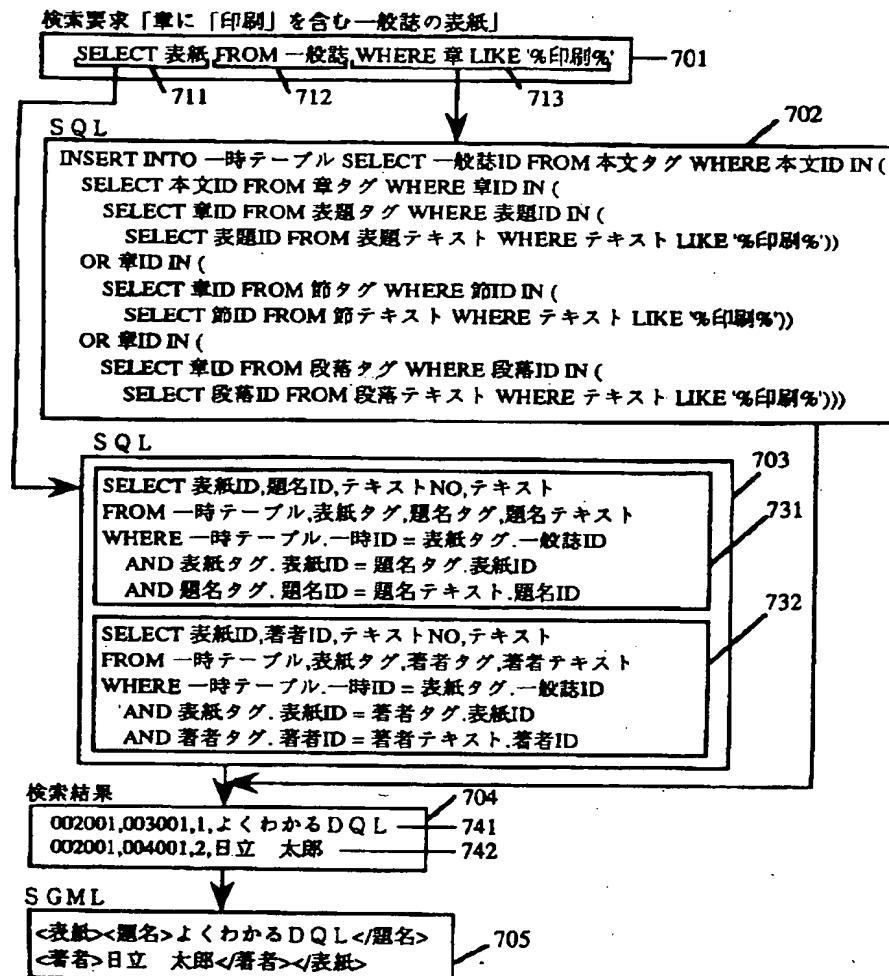
【図7】

図7



【図8】

図8



フロントページの続き

(72) 発明者 山田 清志

東京都江東区新砂一丁目6番27号 株式会
社日立製作所公共情報事業部内

(72) 発明者 根岸 正光

千葉県船橋市前原東6丁目15番23号

(72) 発明者 大山 敬三

千葉県浦安市日の出6番地 ベイシティ浦
安A-1104

(72) 発明者 宮澤 彰

神奈川県横浜市港南区日野南3丁目7番14
号の202

(72) 発明者 芝野 耕司

神奈川県川崎市麻生区高石4丁目12番4号